Towards an Atlas of Computational Learning Theory

Timo Kötzing[1] **and Martin Schirneck**[2]

1 **Hasso Plattner Institute, Potsdam, Germany**
2 **Hasso Plattner Institute, Potsdam, Germany**

Abstract

A major part of our knowledge about Computational Learning stems from comparisons of the learning power of different learning criteria. These comparisons inform about trade-offs between learning restrictions and, more generally, learning settings; furthermore, they inform about what restrictions can be observed without losing learning power.

With this paper we propose that one main focus of future research in Computational Learning should be on a *structured approach* to determine the relations of different learning criteria. In particular, we propose that, for small sets of learning criteria, all pairwise relations should be determined; these relations can then be easily depicted as a *map*, a diagram detailing the relations. Once we have maps for many relevant sets of learning criteria, the collection of these maps is an *Atlas of Computational Learning Theory*, informing at a glance about the landscape of computational learning just as a geographical atlas informs about the earth.

In this paper we work toward this goal by providing three example maps, one pertaining to *partially set-driven* learning, and two pertaining to *strongly monotone* learning. These maps can serve as blueprints for future maps of similar base structure.

1998 ACM Subject Classification I.2.6 Learning

Keywords and phrases computational learning, language learning, partially set-driven learning, strongly monotone learning

Digital Object Identifier 10.4230/LIPIcs.STACS.2016.47

1 Introduction

Computational Learning Theory, also called Inductive Inference, is a branch of (algorithmic) learning theory. This branch analyzes the problem of algorithmically learning a description for a formal language (a computably enumerable subset of the set of natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$) when presented successively the elements of that language. For example, a learner h might be presented more and more even numbers. After each new number, h outputs a description for a language as its conjecture. The learner h might decide to output a program for the set of all multiples of 4, as long as all numbers presented are divisible by 4. Later, when h sees an even number not divisible by 4, it might change this guess to a program for the set of all multiples of 2.

Many criteria for deciding whether a learner h is *successful* on a language L have been proposed in the literature. Gold, in his seminal paper [10], gave a first, simple learning criterion, **Txt**$\mathbf{G}$**Ex**-*learning*[1], where a learner is *successful* iff, on every *text* for L (listing of all and only the elements of L) it eventually stops changing its conjectures, and its final

[1] **Txt** stands for learning from a *text* of positive examples; **G** stands for Gold, who introduced this model, and is used to to indicate full-information learning; **Ex** stands for *explanatory*.

33rd Symposium on Theoretical Aspects of Computer Science (STACS 2016).
Editors: Nicolas Ollinger and Heribert Vollmer; Article No. 47; pp. 47:1–47:13
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany
SYMPOSIUM ON THEORETICAL ASPECTS OF COMPUTER SCIENCE

conjecture is a correct description for the input sequence. Trivially, each single, describable language L has a suitable constant function as a **TxtGEx**-learner (this learner constantly outputs a description for L). Thus, we are interested in analyzing for which *classes of languages* $\mathcal{L}$ there is a *single learner* h learning *each* member of $\mathcal{L}$. This framework is also sometimes known as *language learning in the limit* and has been studied extensively, using a wide range of learning criteria similar to **TxtGEx**-learning (see, for example, the textbook [12]).

Recently, the notion of a learning criterion was formalized in [15] (see Section 2 for the formal notions relevant to this paper). This formalization defines learning criteria as a combination of several components. At the core of any learning criterion is the *interaction operator* which determines what information a learner can get about its target at any stage of the learning process. For example the learner might only see one new datum at a time, only remembering its very previous conjecture (*iterative learning*, **It**), or the learner might get the full information (**G**).

Second, there are many restrictions that can be imposed on learning. For example, **Ex** is the restriction that the learner converges to a single hypothesis, and this hypothesis correctly describes the target language. A relaxation to this is known as *behaviorally correct learning* (**Bc**), which does not require syntactic convergence, but still all but some finite initial hypothesis have to be correct. These restrictions can be combined with other restrictions, such as, for example, *strong monotonicity* (**SMon**), requiring that the languages described by the successive hypotheses to be monotonically non-decreasing. For any given interaction operator β and any learning restriction δ we will use **Txt**$\beta\delta$ to denote the learning criterion employing β as interaction operator and δ as learning restriction (in the setting of learning from text). Note that δ might be the combination, the conjunction, of several learning restrictions; we denote this conjunction of two restriction δ, δ' as $\delta\delta'$.

The main interaction operators from the literature (besides **It** and **G** there are also *set-driven* learning, **Sd**, and *partially set-driven learning*, **Psd**) exhibit a structure: for any two interaction operators β, β', we write $\beta \preceq \beta'$ iff every β-learner can be compiled into an equivalent β'-learner (see Section 2); intuitively, learners can always ignore additional information. In particular, we have the relations

$$\mathbf{Sd} \prec \mathbf{Psd} \prec \mathbf{G} \text{ and } \mathbf{It} \prec \mathbf{G}$$

and no other among these four operators. Note that, for any β, β' with $\beta \preceq \beta'$ and any restriction δ, any class **Txt**$\beta\delta$-learnable class is also **Txt**$\beta'\delta$-learnable: the existence of a β-learner implies the existence of an equivalent β'-learner (see Lemma 1). Note that the converse is not true: While there is no direct way to translate **Sd**-learners into equivalent **It**-learners, it is well known that every **TxtItEx**-learnable class is also **TxtSdEx**-learnable [14]. In other words, some comparisons of learning power of different learning criteria are *trivial* (following directly from very basic relations of the interaction operators), while others are *contingent* on the setting.

A similar observation holds for learning restrictions. These restrictions can be compared with $\subseteq$ (since we formally define them as sets of pairs for which the restriction holds); for example, we have $\mathbf{Ex} \subset \mathbf{Bc}$ in the sense that any sequence of conjectures which correctly **Ex**-identifies a target from a text T is also a sequence of conjectures which correctly **Bc**-identifies the same target from the same text T (but not vice versa). This again gives that some comparisons of learning power of different learning criteria are *trivial* (following directly from the $\subseteq$ relation on the restrictions), while others are *contingent* on the setting.

This observation holds for all parts of the learning criterion: a monotone change in a single component leads to a monotone change of the learning power of the learning criterion

(see Lemma 1 for the formal statement). In other words, for any set of learning criteria, the rough structure is visible from trivial inclusions, detailed structure requires specific analysis. Just as the exploration of a geographical landscape might begin with drawing a rough map of the area and then go explore the different parts in detail, a researcher comparing the learning power of different learning criteria might proceed by drawing the trivial inclusions among these criteria as a kind of "backbone" and then determine what further contingent relations hold. In this way the researcher draws a "map" of the collection of learning criteria.

For a full characterization it would be desirable to have a map of all (important) learning criteria. As this would require to determine the pairwise relations of several hundred learning criteria (using all possible combinations of different possible components of learning criteria), this is not easily feasible and more of a long term goal (just as the complete exploration of the earth is not feasible in one go). Furthermore, a large map of all criteria might not be very easy to understand; just as is done for the earth, giving a collection of maps, each giving details for some specific part, gives a much better idea. Thus, we want to propose that an important goal in Computational Learning Theory is *to give an atlas of insightful maps of learning criteria.*

The question which maps are insightful is of course the crucial and difficult part. The literature has given several examples; we proceed by giving three main kinds of maps that we propose to be insightful.

1.1 Partially Set-Driven Learning

A wealth of learning criteria can be derived from **TxtGEx**-learning by adding restrictions on the intermediate conjectures and how they should relate to each other and the data. For example, one could require that a conjecture which is consistent with the data must not be changed; this is known as *conservative* learning (**Conv**, [1]). Additionally to conservative learning, the following learning restrictions are considered frequently in the literature.

In *cautious* learning (**Caut**, [19]) the learner is not allowed to ever give a conjecture for a strict subset of a previously conjectured set. In *non-U-shaped* learning (**NU**, [2]) a learner may never *semantically* abandon a correct conjecture; in *strongly non-U-shaped* learning (**SNU**, [7]) not even syntactic changes are allowed after giving a correct conjecture. In *decisive* learning (**Dec**, [19]), a learner may never (semantically) return to a *semantically* abandoned conjecture; in *strongly decisive* learning (**SDec**, [16]) the learner may not even (semantically) return to *syntactically* abandoned conjectures. Finally, a number of monotonicity requirements are studied ([13, 24, 18]): in *strongly monotone* learning (**SMon**) the conjectured sets may only grow; in *monotone* learning (**Mon**) only incorrect data may be removed; and in *weakly monotone* learning (**WMon**) the conjectured set may only grow while it is consistent. A common property of these restrictions is *delayability* (see Definition 2).

Recently, [17] gave the map of **TxtG**δ**Ex** for all the learning restriction δ given above (plus **T**, denoting no restriction). The same paper also gave the map for **Sd** in place of **G**, while [11] gave the map for **It** in place of **G**. Thus, the only main interaction operator still missing is **Psd**, for which we give the map in this paper (depicted in Figure 1, see Section 3 for all relevant theorems). A solid black line indicates a trivial inclusion (the lower criterion is included in the higher); a dashed black line indicates an inclusion which is not trivial. A gray box around criteria indicates equality of (learning power of) these criteria. We will use this way of graphical representation for every map in this paper.

Thus, the solid black lines are the backbone of these learning criteria. It is interesting to note that the structure is exactly like for the interaction operator **G**, even though most of the proofs do not carry over, and **Psd**-learning is in general weaker than its **G**-variants (as,

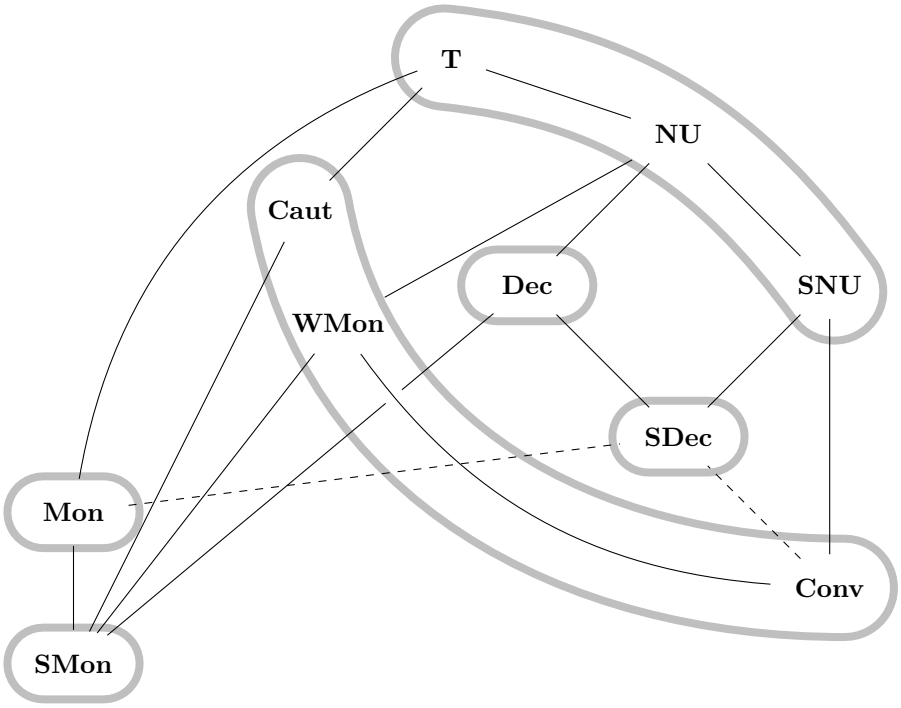

Figure 1 Relation of [**TxtPsd**δ**Ex**] for various learning restrictions δ. The backbone is given by the black solid lines (trivial inclusions bottom-to-top). The only previously known relations are the collapse of **T**, **NU** and **SNU**, as well as that **SMon** and **Mon** are each on their own.

for example, in the case of strongly monotone learning, see Section 1.2). Typically the proofs can use some ideas from the **G**-case, but need to add some specific ideas. In this context we developed some additional methods, such as a normal form for partially set-driven learners (see Lemma 3). Also, we showed that conservative, weakly monotone and cautious learning lie in between two other restrictions, which are then shown to be equivalent (see Theorem 6). This gives another interesting characterization of the important restriction of conservative learning.

1.2 Strongly Montone Learning and Interaction Operators

When comparing different interaction operators, it comes in handy to have a maximum (with respect to $\preceq$) and also a minimum. The maximum (for the four main operators) is **G**. As minimum we introduce *confluently iterative learning* (**CflIt**). It requires a learner to be just like an iterative learner, but with the added restriction of being confluent, that is, when a sequence of inputs is given in different order or quantity, the same output is produced by the learner. Intuitively, the learner has to be set-driven (thus **CflIt** $\preceq$ **Sd** and **CflIt** $\preceq$ **It**).

Consider now the restriction of strongly monotone learning (**SMon**). We can either require that the learner has to be strongly monotone only on relevant inputs (relevant to the current class of languages to be learned) or generally. This latter version is denoted τ(**SMon**) and written as a prefix to the learning criterion. Note that this requires the learner to be total (usually any element from $\mathcal{P}$, the set of all partial computable functions, can be used as a learner, as long as the learner produces output on relevant inputs). As totality of the learner can impede the learner in itself, it is interesting to compare the two different versions of a restrictions (on relevant or on all inputs) also with the version where the restriction is

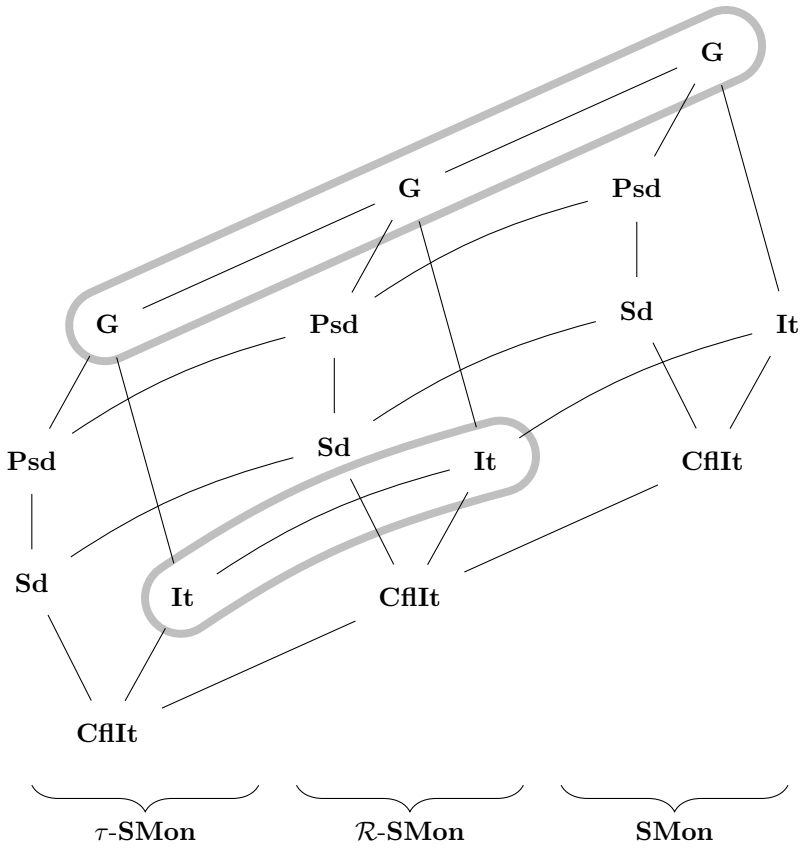

Figure 2 Relation of $[\tau(\mathbf{SMon})\mathbf{Txt}\beta\mathbf{Ex}]$, $[\mathcal{R}\mathbf{Txt}\beta\mathbf{SMonEx}]$ and $[\mathbf{Txt}\beta\mathbf{SMonEx}]$ for various interaction operators β. The backbone is given by the black solid lines (trivial inclusions bottom-to-top). The only previously known relation is $[\mathcal{R}\mathbf{TxtGSMonEx}] = [\mathbf{TxtGSMonEx}]$.

only on relevant inputs, but the learner is required to be total. We denote the restriction of totality by $\mathcal{R}$ (the symbol for the set of total computable functions) as a prefix.

Naturally we have that $\tau(\mathbf{SMon})\mathbf{Txt}\beta\mathbf{Ex}$ is trivially weaker in learning power than $\mathcal{R}\mathbf{Txt}\beta\mathbf{SMonEx}$, which is in turn trivially weaker in learning power than $\mathbf{Txt}\beta\mathbf{SMonEx}$. In this paper we give the map for these learning criteria (depicted in Figure 2, see Section 4 for all relevant theorems.

Again the solid black lines give the backbone. When replacing **SMon** with any other learning restriction, this backbone would stay the same. Noteworthy is the collapse of the three different criteria involving **G**, which has the same underlying idea as the collapse of the two criteria featuring **It**. As we can see, any strongly monotone **G**-learner can be assumed to be not only total but also strongly monotone on arbitrary inputs; this does not hold for the other interaction operators: the proof for **G** exploits first a standard delaying trick to make sure that the learner is total, and then make use of the knowledge of prior hypotheses to make sure that learner proceeds strongly monotonically.

Furthermore we also give the map for the criteria for the corresponding learning criteria with **Bc** in place of **Ex**. This map is trivial, as all learning criteria $\tau(\mathbf{SMon})\mathbf{Txt}\beta\mathbf{Bc}$, $\mathcal{R}\mathbf{Txt}\beta\mathbf{SMonBc}$ and $\mathbf{Txt}\beta\mathbf{SMonBc}$ for the five different β are equivalent in learning power.

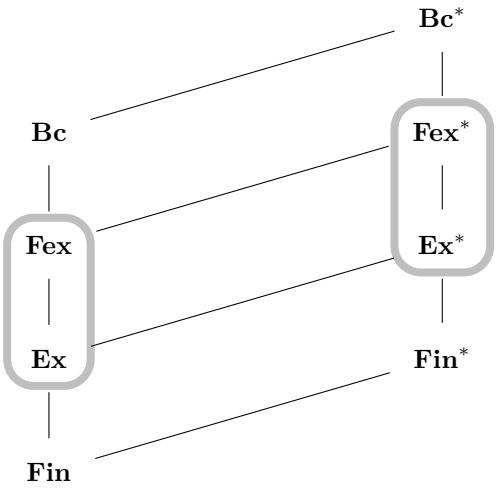

Figure 3 Relation of [**TxtGSMon**δ] for various δ. The backbone is given by the black solid lines (trivial inclusions bottom-to-top).

1.3 Strongly Monotone Learning and Convergence Criteria

In Section 1.1 we saw many different learning restrictions which are typically combined with **Ex**. As an alternative to **Ex** we saw **Bc**. Another alternative is **Fex**, allowing any *finite* number of limit conjectures, naturally in between **Ex** and **Bc**. Even more restrictive than **Ex** is **Fin**, allowing only one hypothesis different from ? overall (this is *finite learning*). Each of these four convergence restrictions can be relaxed by counting hypotheses for finite variants of the target also as correct hypotheses; this is denoted by an asterisk as suffix (e.g., **Ex***).

We give the map for the learning criteria **TxtGSMon**δ for the eight different choices of convergence criteria (depicted in Figure 3, see Section 5 for all relevant theorems).
As you can see, the **Ex**-variants and the corresponding **Fex**-variants collapse in the case of strongly monotone full-information learning. All other learning criteria stay separated, and no further inclusions exist.

1.4 Conclusions

We have given three different maps considering very different kinds of learning criteria. We believe that such maps increase our general understanding of learning and give a better picture of how different criteria relate. While establishing these maps we typically get structural insights, which can easily be explained with reference to the map. Most crucially, since the main result of the research is a map, it is very easy to communicate the results; especially other researchers who know the general backbone will be able to take in the results effortlessly. Furthermore, this approach points out gaps in our knowledge about learning criteria and thus focuses research efforts: maps which leave open problems are typically of much lesser value, since they only provide a partial picture.

We believe that giving similar maps will drive future research in an important direction: the development of an *Atlas of Computational Learning Theory*.

Next, in Section 2, we give a formal definition of learning criteria. Sections 3 to 5 conclude the paper by giving the formal theorems establishing the maps presented above. All proofs are omitted due to space restrictions; the main contribution of this paper lies elsewhere.

2 Learning Criteria

In this section we formally introduce our setting of learning in the limit and associated learning criteria. We follow the system in [15] in defining learning criteria. For background on computability, see [21]. In particular, we let W_e denote the recursively enumerable (r.e.) set enumerated by the program with (coded) number e. Thus, we can interpret a natural numbers e as hypothesis for the set W_e. We denote the set of all r.e. sets by $\mathcal{E}$.

A *learner* is a partial computable function $h \in \mathcal{P}$. We allow learners to output ? to denote that no conjecture is made yet. A *language* is an r.e. set $L \in \mathcal{E}$ of natural numbers. Any total function $T\colon \mathbb{N} \to \mathbb{N} \cup \{\#\}$ is a *text*, the collection of all texts is **Txt**. For any text (or sequence) T, we let $\mathrm{content}(T) = \mathrm{range}(T)\backslash\{\#\}$. For any given language L, a *text for* L is a text T such that $\mathrm{content}(T) = L$. The set of all texts for some language L is denoted $\mathbf{Txt}(L)$. For a given text $T \in \mathbf{Txt}$ and any n, we use $T[n]$ to denote the sequence $(T(0), \ldots, T(n-1))$ (the empty sequence λ when $n = 0$). Initial parts of this kind is what learners usually get as information.

An *interaction operator* is an operator β taking as arguments a function h (the learner) and a text T, and outputs a (possibly partial) function p. We call p the *learning sequence* (or *sequence of hypotheses*) of h given T. We define the interaction operators **G** (*Gold-style* or *full-information learning* [10]), **Psd** (*partially set-driven learning*, [22]), **Sd** (*set-driven learning*, [23]) and **It** (*iterative learning*, [23]) as follows. For all $h \in \mathcal{P}$, texts T and all i,

$$\begin{aligned}
\mathbf{G}(h,T)(i) &= h(T[i]);\\
\mathbf{Psd}(h,T)(i) &= h(\mathrm{content}(T[i]),i);\\
\mathbf{Sd}(h,T)(i) &= h(\mathrm{content}(T[i]));\\
\mathbf{It}(h,T)(i) &= \begin{cases} h(\lambda), & \text{if } i=0;\\ h(\mathbf{It}(h,T)(i-1),T(i-1)), & \text{otherwise.}\end{cases}
\end{aligned}$$

In set-driven learning, the learner has access to the set of all previous data, but not to the full sequence as in **G**-learning. In partially set-driven learning, the learner has the set of data and the current iteration number. **Psd**-learning is sometimes also called *rearrangement-independent learning* [4]. In iterative learning, the learner can access its last hypothesis as well as the most recent input data. Hereby, $h(\lambda)$ denotes the initial hypothesis of learner h. For two interaction operators β, β', we say *β-learners can be translated into β'-learners*, written $\beta \preceq \beta'$, if, for every learner h, there is some learner h' such that, for arbitrary texts T, the resulting sequence of hypotheses of h working on T is the same as that of h', i.e. $\forall T \in \mathbf{Txt}\colon \beta(h,T) = \beta'(h',T)$. For example, an **Sd**-learner can be translated into an **Psd**-learner by simply ignoring the additional information of the number of the current iteration. Clearly, all learners investigated in this paper can be translated into **G**-learners. For any β-learner h such that $\beta \preceq \mathbf{G}$, we let h^* (the *starred learner*) denote the **G**-learner to simulate h. A learner h is said to be *confluently iterative* just in case it is both set-driven and iterative. That is, its starred learner h^* satisfies the following two conditions. For any two finite sequences σ, τ and natural numbers x, we have $\mathrm{content}(\sigma) = \mathrm{content}(\tau) \Rightarrow h^*(\sigma) = h^*(\tau)$ and $h^*(\sigma) = h^*(\tau) \Rightarrow h^*(\sigma \diamond x) = h^*(\tau \diamond x)$. The interaction operator associated with confluent iterativeness is denoted **CflIt**, it is a $\preceq$-lower bound for both **Sd** and **It**.

Successful learning requires the learner to observe certain restrictions, for example convergence to a correct index. A *learning restriction* is a predicate δ on a learning sequence and a text. We give the important example of explanatory learning (**Ex**, [10]) defined such

that, for all sequences of hypotheses p and all texts T,

$$\mathbf{Ex}(p,T) \Leftrightarrow p \text{ total } \wedge [\exists n_0 \colon (\forall n \geq n_0 : p(n) = p(n_0)) \wedge W_{p(n_0)} = \text{content}(T)].$$

There are several other success criteria under investigation in this work, most notably in Section 5. We always require successful learning sequences p to be total, as in **Ex**-learning. These success criteria, as well as other learning restrictions discussed in Section 1, are given as follows.

$$\begin{aligned}
\mathbf{Ex}^*(p,T) &\Leftrightarrow \exists n_0 \colon (\forall n \geq n_0 : p(n) = p(n_0)) \wedge W_{p(n_0)} =^* \text{content}(T);\\
\mathbf{Fex}(p,T) &\Leftrightarrow \exists D\ \exists n_0 \colon (\forall n \geq n_0 \colon p(n) \in D) \wedge (\forall e \in D \colon W_e = \text{content}(T));\\
\mathbf{Fex}^*(p,T) &\Leftrightarrow \exists D\ \exists n_0 \colon (\forall n \geq n_0 \colon p(n) \in D) \wedge (\forall e \in D \colon W_e =^* \text{content}(T));\\
\mathbf{Bc}(p,T) &\Leftrightarrow \exists n_0\ \forall n \geq n_0 \colon W_{p(n)} = \text{content}(T);\\
\mathbf{Bc}^*(p,T) &\Leftrightarrow \exists n_0\ \forall n \geq n_0 \colon W_{p(n)} =^* \text{content}(T);\\
\mathbf{Fin}(p,T) &\Leftrightarrow \exists n_0 \colon (\forall n < n_0 \colon p(n) = ?) \wedge W_{p(n_0)} = \text{content}(T);\\
\mathbf{Fin}^*(p,T) &\Leftrightarrow \exists n_0 \colon (\forall n < n_0 \colon p(n) = ?) \wedge W_{p(n_0)} =^* \text{content}(T).\\
\mathbf{Conv}(p,T) &\Leftrightarrow \forall i : \text{content}(T[i+1]) \subseteq W_{p(i)} \Rightarrow p(i) = p(i+1);\\
\mathbf{Caut}(p,T) &\Leftrightarrow \forall i,j : W_{p(i)} \subset W_{p(j)} \Rightarrow i < j;\\
\mathbf{NU}(p,T) &\Leftrightarrow \forall i,j,k : i \leq j \leq k \ \wedge\ W_{p(i)} = W_{p(k)} = \text{content}(T) \Rightarrow W_{p(j)} = W_{p(i)};\\
\mathbf{Dec}(p,T) &\Leftrightarrow \forall i,j,k : i \leq j \leq k \ \wedge\ W_{p(i)} = W_{p(k)} \Rightarrow W_{p(j)} = W_{p(i)};\\
\mathbf{SNU}(p,T) &\Leftrightarrow \forall i,j,k : i \leq j \leq k \ \wedge\ W_{p(i)} = W_{p(k)} = \text{content}(T) \Rightarrow p(j) = p(i);\\
\mathbf{SDec}(p,T) &\Leftrightarrow \forall i,j,k : i \leq j \leq k \ \wedge\ W_{p(i)} = W_{p(k)} \Rightarrow p(j) = p(i);\\
\mathbf{SMon}(p,T) &\Leftrightarrow \forall i,j : i < j \Rightarrow W_{p(i)} \subseteq W_{p(j)};\\
\mathbf{Mon}(p,T) &\Leftrightarrow \forall i,j : i < j \Rightarrow W_{p(i)} \cap \text{content}(T) \subseteq W_{p(j)} \cap \text{content}(T);\\
\mathbf{WMon}(p,T) &\Leftrightarrow \forall i,j : i < j \wedge \text{content}(T[j]) \subseteq W_{p(i)} \Rightarrow W_{p(i)} \subseteq W_{p(j)}.
\end{aligned}$$

We combine any two learning restrictions δ and δ' by intersecting them; we denote this by juxtaposition. With $\mathbf{T}$ we denote the restriction which is always true (no restriction).

Now a *learning criterion* is a tuple $(\alpha, \mathcal{C}, \beta, \delta)$, where $\mathcal{C}$ is a set of learners (the admissible learners; typically $\mathcal{P}$ or $\mathcal{R}$), β is an interaction operator and α, δ are learning restrictions; we write $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$ to denote the learning criterion, omitting $\mathcal{C}$ in case of $\mathcal{C} = \mathcal{P}$ and the restriction in case it equals $\mathbf{T}$. We say that a learner $h \in \mathcal{C}$ $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$*-learns* a language L iff, on *arbitrary* texts $T \in \mathbf{Txt}$, $\alpha(\beta(h,T),T)$ and, for all texts $T \in \mathbf{Txt}(L)$, $\delta(\beta(h,T),T)$. The set of languages $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$-learned by $h \in \mathcal{C}$ is denoted by $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta(h)$. We write $[\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta]$ to denote the set of all $\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta$-learnable classes.

Throughout this paper we are mostly concerned in showing separations between learning criteria. The reason is, that most of the inclusions are trivial in the sense that almost all of them follow from the next lemma. A formal proof can be found in [6].

▶ **Lemma 1.** *Let $\alpha \subseteq \alpha', \delta \subseteq \delta'$ be learning restrictions, $\mathcal{C} \subseteq \mathcal{C}'$ classes of admissible learners and $\beta \preceq \beta'$ two interaction operators. Then we have $[\tau(\alpha)\mathcal{C}\mathbf{Txt}\beta\delta] \subseteq [\tau(\alpha')\mathcal{C}'\mathbf{Txt}\beta'\delta']$.*

3 Delayable Partially Set-driven Language Learning

In this section we will investigate delayable **Psd**-learning. First, we establish a normal form for **Psd**-learners. Consider pairs (D,t) and (D',t') consisting of finite sets D, D' and numbers t, t', we write $(D,t) \to (D',t')$ just in case $t \leq t'$ and there is a text T such that

content$(T[t]) = D$ and content$(T[t']) = D'$. Note that $(D,t) \to (D',t')$ implies $D \subseteq D'$. Let h be some learner and L a language. We call a pair (D,t), with $D \subseteq L$ and $t \geq |D|$, such that $W_{h(D,t)} = L$ and for all (D',t') such that $D' \subseteq L$ and $(D,t) \to (D',t')$, $h(D,t) = h(D',t')$, a *locking information* for learner h on L. If we use interaction operator **G** instead a locking information is commonly referred to as a *locking sequence*. It is well known that, if h **Ex**-learns L, then *every* such pair can be extended to some locking information [4, 6]. Moreover, we call h *strongly locking* if, for each language $L \in \mathbf{TxtPsdEx}(h)$ and every text $T \in \mathbf{Txt}(L)$ for L, there is an n such that $(\text{content}(T[n]), n)$ serves as a locking information for h on L. We call learner h *order-independent* if, for all languages $L \in \mathbf{TxtPsdEx}(h)$ and any two texts $T, T' \in \mathbf{Txt}(L)$ for L, we have $\lim_{n\to\infty} \mathbf{Psd}(h,T)(n) = \lim_{n\to\infty} \mathbf{Psd}(h,T')(n)$. That is, h's final hypothesis only depends on the language L, not on the particular order in which its elements are presented in the text. We now turn to the notion of delayable learning.

▶ **Definition 2.** Let $\vec{R}$ be the set of all unbounded non-decreasing functions $r\colon \mathbb{N} \to \mathbb{N}$, i.e. for all m we have $\forall^\infty n : r(n) \geq m$. We call a learning restriction δ *delayable* if, for all texts T' and T with content(T') = content(T), all infinite sequences p and all $r \in \vec{R}$, if $(p,T') \in \delta$ and $\forall n : \text{content}(T'[r(n)]) \subseteq \text{content}(T[n])$, then $(p \circ r, T) \in \delta$.

Intuitively, as long as the learner has *at least as much* data as was used for a given conjecture, then this conjecture is permissible. The intersection of two delayable learning criteria is again delayable. *All* learning restrictions considered in this paper, including success criteria and **T**, are delayable. We now get a normal form for delayable **Psd**-learner.

▶ **Lemma 3.** *Let restrictions α, δ be delayable and $\mathcal{C} \in \{\mathcal{P}, \mathcal{R}\}$. Then $\tau(\alpha)\mathcal{C}\mathbf{TxtPsd}\delta$ allows for strongly locking and order-independent learning (simultaneously).*

We now proceed to prove the connections shown in Figure 1. The following theorem translates what is known about the respective criteria of delayable **G**-learning into the **Psd**-setting. The first part has already been proven by Case and Kötzing [6]. Furthermore, the separations in the second part (in the full-information case) are due to Baliga et al. [2], Kötzing and Palenta [17] as well as Osherson et al. [20]. For a collection of known separations see [17].

▶ **Theorem 4.** *Among the investigated criteria non-U-shapedness and strong non-U-shapedness are the only ones equally powerful to unrestricted* **Psd***-learning when paired with success criterion* **Ex**. *This is, we have:*

1. $[\mathbf{TxtPsdSNUEx}] = [\mathbf{TxtPsdNUEx}] = [\mathbf{TxtPsdEx}]$.
2. *For* $\delta \in \{\mathbf{Conv}, \mathbf{Caut}, \mathbf{Dec}, \mathbf{SDec}, \mathbf{Mon}, \mathbf{WMon}, \mathbf{SMon}\}$, $[\mathbf{TxtPsd}\delta\mathbf{Ex}] \subset [\mathbf{TxtPsdEx}]$.

The separation of decisive and strongly decisive learning follows just as for **G**, see [17].

▶ **Theorem 5.** *We have* $[\mathbf{TxtPsdSDecEx}] \subset [\mathbf{TxtPsdDecEx}]$.

For our work on conservative, weakly monotone and cautious learning, we first give two more learning restrictions. One of them, *witness-based* learning (**Wb**), is more restrictive than all three of conservative, weakly monotone and cautious learning; the other, *target-cautious learning* ($\mathbf{Caut_{Tar}}$), is less restrictive. That way those three learning restrictions are "sandwiched" between witness-based and target-cautious learning. We will then show that witness-based and target-cautious learning have equal learning power in the setting of partially set-driven learning, showing all three sandwiched restrictions to be equivalent.

We start by giving the definition of witness-based learning.

$$\mathbf{Wb}(p,T) \Leftrightarrow \forall i,j : (\exists k : i < k \leq j \wedge p(i) \neq p(k)) \Rightarrow (\text{content}(T[j]) \cap W_{p(j)}) \setminus W_{p(i)} \neq \emptyset.$$

Intuitively, any mind change has to be *witnessed* by some datum which was not included before, but is included now; this witness *justifies* the mind change.

Target-cautious learning was introduced in [17] to study the notion of cautious learning in more detail.

$$\mathbf{Caut_{Tar}}(p,T) \Leftrightarrow \forall i : \neg(\text{content}(T) \subset W_{p(i)})$$

Intuitively, the learner may never conjecture a superset of the actual target language; in other words, it is required to be cautious, but only with respect to the target.

We now get to the theorem establishing the equivalence of the conservative, weakly monotone and cautious learning.

▶ **Theorem 6.** *The following learning criteria are equivalent:* $\mathbf{TxtPsdCaut_{Tar}Ex}$*;* **TxtPsdCautEx***;* **TxtPsdConvEx***;* **TxtPsdWMonEx***;* **TxtPsdWbEx***.*

The following theorem solely reformulates a result well-known in **G**-learning in terms of **Psd**-learning. The first part uses a standard proof, see [12] for example. The second part has already been shown (for **G**-learning) by Kötzing and Palenta in [17], in turn based on a technique presented in [20].

▶ **Theorem 7.** *Monotone and weakly monotone* **Psd***-learning is incomparable. In particular, we have*

1. [**TxtSdWMonEx**]\[**TxtPsdMonEx**] $\neq \emptyset$*;*
2. [**TxtPsdMonSDecEx**]\[**TxtPsdWMonEx**] $\neq \emptyset$*;*

▶ **Corollary 8.** *For each learning restriction* $\delta \in \{\mathbf{Caut}, \mathbf{Caut_{Tar}}, \mathbf{Conv}, \mathbf{Wb}, \mathbf{WMon}\}$*, we have* [**TxtPsd**$\delta$**Ex**] $\subset$ [**TxtPsdSDecEx**]*.*

The upcoming lemma is based on an theorem due to Baliga et al. [2] stating that concept classes containing the set $\mathbb{N}$ of all natural numbers, if they are inferable at all, are decisively learnable. Kötzing and Palenta extended it to comprise strong decisiveness [17]. We show that the result still holds when restricted to **Psd**-learnable classes.

▶ **Lemma 9.** *Let* $\mathcal{L}$ *be a class of languages with* $\mathbb{N} \in \mathcal{L}$*. If class* $\mathcal{L}$ *is identifiable by a* **Psd***-learner at all, then it can in fact be so learned strongly decisive.*

We can now use Lemma 9 to show that every monotonically learnable class can be learned strongly decisively.

▶ **Theorem 10.** *Any monotonically* **Psd***-learnable class of languages can be so learned strongly decisively, while the converse does not hold. Thus, we have* [**TxtPsdMonEx**] $\subset$ [**TxtPsdSDecEx**]*.*

4 Strongly Monotone Language Learning

In this section we prove Figure 2 (in Section 4.1) as well as the equivalence of all corresponding learning criteria with **Bc** in place of **Ex** (in Section 4.2).

4.1 Ex-Learning

In this section we discuss strongly monotone language learning in the context of explanatory (**Ex**) convergence. In particular, we will prove the diagram shown in Figure 2. For any delayable learning restriction δ we can, without loss of generality, assume every **G**-learner with respect to δ to be *total* [17]. In particular, this holds for $\delta = \mathbf{SMonEx}$ (and **SMonBc** as well). Thus, we get a first connection between the investigated criteria, namely $[\mathcal{R}\mathbf{TxtGSMonEx}] = [\mathbf{TxtGSMonEx}]$. Interestingly enough, at least for strongly monotone explanatory learning, **G** is the *only* interaction operator for which this relation holds, as shown in the next theorem.

▶ **Theorem 11.** *For each interaction operator* $\beta \in \{\mathbf{CflIt}, \mathbf{It}, \mathbf{Sd}, \mathbf{Psd}\}$*, we have* $[\mathbf{TxtCflItSMonEx}] \setminus [\mathcal{R}\mathbf{Txt}\beta\mathbf{SMonEx}] \neq \emptyset$.

Recall that we use $\tau(\alpha)$ to denote that a learner observes learning restriction α on *arbitrary* texts, even on those for languages it cannot identify. In the discussion of $\tau(\mathbf{SMon})$-learners we can distinguish two major groups: On one hand, we have **G**- as well as **It**-learners which can be transposed into τ-learners without loss of learning power. On the other hand, there is the group of **Psd**-, **Sd**- and **CflIt**-learners for which their total variants are strictly more powerful than their globally strongly monotone matches. The main property discriminating these two groups is whether the learner has access to its previous conjecture.

▶ **Theorem 12.** *Total full-information and iterative* **SMon***-learning can w.l.o.g. be done by a learner being strongly monotone on arbitrary texts. Thus, we get the following equalities.*

1. $[\tau(\mathbf{SMon})\mathbf{TxtGEx}] = [\mathcal{R}\mathbf{TxtGSMonEx}] = [\mathbf{TxtGSMonEx}]$;
2. $[\tau(\mathbf{SMon})\mathbf{TxtItEx}] = [\mathcal{R}\mathbf{TxtItSMonEx}]$.

The target classes identifiable by τ-learners drawn from the aforementioned second group (interaction operators **Psd**, **Sd** and **CflIt**) share an interesting common trait in terms of recursiveness. It is stated in the following lemma. The technique used in its proof resembles that of a well-known proposition regarding globally *consistent* learning, cf. [1] and [12, Proposition 5.6].

▶ **Lemma 13.** *If a class* $\mathcal{L}$ *is identifiable by a* **Psd***-learner being globally strongly monotone and if* $\mathcal{L}$ *comprises at least all singleton sets, then* $\mathcal{L}$ *is a collection of recursive languages.*

▶ **Theorem 14.** *There is a class of languages identifiable by a total* **CflIt***-learner which cannot be learned by a globally strongly monotone* **Psd***-learner. That is, for all interaction operators* $\beta \in \{\mathbf{CflIt}, \mathbf{Sd}, \mathbf{Psd}\}$*, we have* $[\mathcal{R}\mathbf{TxtCflItSMonEx}] \setminus [\tau(\mathbf{SMon})\mathbf{Txt}\beta\mathbf{Ex}] \neq \emptyset$.

▶ **Theorem 15.** *There is a class of languages identifiable by a total strongly monotone iterative learner which cannot be so learned partially set-driven, this is,* $[\mathcal{R}\mathbf{TxtItSMonEx}] \setminus [\mathbf{TxtPsdSMonEx}] \neq \emptyset$.

▶ **Corollary 16.** **Psd***-learning with respect to* **SMon** *is strictly less powerful than its full-information counterpart. Hence, we have* $[\mathbf{TxtPsdSMonEx}] \subset [\mathbf{TxtGSMonEx}]$.

This corollary stands in sharp contrast to the abilities of partially set-driven functions in *unrestricted* language learning. A famous result due to Fulk [9] states that *any* class of languages, which can be learned at all, can be inferred by a total **Psd**-learner. To our knowledge, **SMon** is the first learning restriction in literature for which no equivalent of Fulk's Theorem holds.

▶ **Theorem 17.** *We have* $[\tau(\mathbf{SMon})\mathbf{TxtSdEx}] \setminus [\mathbf{TxtItSMonEx}] \neq \emptyset$*, i.e. there is a class of languages identifiable by a globally strongly monotone* **Sd***-learner which cannot be identified by any* **It***-learner.*

The last question to be covered in this section is that of the relation among the learners *within* the second group, namely, that between **Psd**- and **Sd**-learners.

▶ **Theorem 18.** *For all variants,* **Psd***-learners are strictly more powerful than their set-driven analogues. Particularly, we have* $[\tau(\mathbf{SMon})\mathbf{TxtPsdEx}]\setminus[\mathbf{TxtSdSMonEx}] \neq \emptyset$.

4.2 Bc-Learning

In this section we turn the discussion to behaviorally correct (**Bc**) language learning. It becomes apparent that all criteria in this setting possess the same learning power. We establish this in one step by showing that full-information **Bc**-learning with respect to **SMon** can be done confluently iteratively being globally strongly monotone. We conclude this section by proving that strongly monotone **Bc**-learning is strictly more powerful than strongly monotone **Ex**-learning.

▶ **Theorem 19.** *We have* $[\tau(\mathbf{SMon})\mathbf{TxtCflItBc}] = [\mathbf{TxtGSMonBc}]$*, i.e. every class of languages which is* **TxtGSMonBc***-identifiable can be learned by a* **CflIt***-learner being strongly monotone on arbitrary texts.*

▶ **Theorem 20.** *Strongly monotone* **Bc***-learning is strictly more powerful than its explanatory counterpart. So we have* $[\mathbf{TxtGSMonEx}] \subset [\mathbf{TxtGSMonBc}]$*. Even stronger, we have* $[\mathbf{TxtGSMonBc}] \setminus [\mathbf{TxtGEx}] \neq \emptyset$.

5 Anomalous and Vacillatory Language Learning

In this section we examine the behavior of different success criteria for learning when paired with the requirement of strong monotonicity. A result in the field of function learning states one does *not* gain additional learning power in allowing the learner to vacillate between finitely many correct hypotheses in the limit [3, 8]. However, in (unrestricted) language identification, **Fex**-learners can indeed infer strictly more classes of languages [5]. We begin our analysis in proving that, when paired with **SMon**, this advantage vanishes once again.

▶ **Theorem 21.** *For strongly monotone language learning, vacillating among finitely many hypotheses does not increase learning power. Thus, we have*
1. $[\mathbf{TxtGSMonEx}] = [\mathbf{TxtGSMonFex}]$*;*
2. $[\mathbf{TxtGSMonEx}^*] = [\mathbf{TxtGSMonFex}^*]$.

We conclude with two theorems establishing the separations given in Figure 3. Note that for finite learning (**Fin**) any learner is strongly monotone as it outputs only a single hypothesis besides "?", hence, $[\mathbf{TxtGSMonFin}^*] = [\mathbf{TxtGFin}^*]$.

▶ **Theorem 22.** *There is a class of languages which can only be inferred if the learner is allowed to make finite error. Thus, we have* $[\mathbf{TxtGSMonFin}^*]\setminus[\mathbf{TxtGSMonBc}] \neq \emptyset$*. Even stronger, we have* $[\mathbf{TxtGSMonFin}^*]\setminus[\mathbf{TxtGBc}] \neq \emptyset$

▶ **Theorem 23.** *The following two separations hold for anomalous and behaviorally correct strongly monotone language learning:*
1. $[\mathbf{TxtGSMonEx}]\setminus[\mathbf{TxtGSMonFin}^*] \neq \emptyset$*;*
2. $[\mathbf{TxtGSMonBc}]\setminus[\mathbf{TxtGSMonEx}^*] \neq \emptyset$.

References

1 D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135, 1980.
2 G. Baliga, J. Case, W. Merkle, F. Stephan, and W. Wiehagen. When unlearning helps. *Information and Computation*, 206:694–709, 2008.
3 J. Bārzdiņš and K. Podnieks. The theory of inductive inference. In *Mathematical Foundations of Computer Science*, 1973.
4 L. Blum and M. Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155, 1975.
5 J. Case. The power of vacillation in language learning. *SIAM Journal on Computing*, 28(6):1941–1969, 1999.
6 J. Case and T. Kötzing. Strongly non-U-shaped learning results by general techniques. In *Proc. of COLT (Conference on Learning Theory)*, pages 181–193, 2010.
7 J. Case and S. Moelius. Optimal language learning from positive data. *Information and Computation*, 209:1293–1311, 2011.
8 J. Case and C. Smith. Comparison of identification criteria for machine inductive inference. *Theoretical Computer Science*, 25:193–220, 1983.
9 M. Fulk. Prudence and other conditions on formal language learning. *Information and Computation*, 85:1–11, 1990.
10 E. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.
11 S. Jain, T. Kötzing, and F. Stephan. On the role of update constraints and text-types in iterative learning. In *Proc. of ALT (Algorithmic Learning Theory)*, 2014.
12 S. Jain, D. Osherson, J. Royer, and A. Sharma. *Systems that Learn: An Introduction to Learning Theory*. MIT Press, Cambridge, MA, second edition, 1999.
13 K. Jantke. Monotonic and non-monotonic inductive inference of functions and patterns. In *Proc. of Nonmonotonic and Inductive Logic*, pages 161–177, 1991.
14 E. Kinber and F. Stephan. Language learning from texts: Mind changes, limited memory and monotonicity. *Information and Computation*, 123:224–241, 1995.
15 T. Kötzing. *Abstraction and Complexity in Computational Learning in the Limit*. PhD thesis, University of Delaware, 2009. Available online at http://pqdtopen.proquest.com/#viewpdf?dispub=3373055.
16 T. Kötzing. A solution to Wiehagen's thesis. In *Proc. of STACS (Symposium on Theoretical Aspects of Computer Science)*, pages 494–505, 2014.
17 T. Kötzing and R. Palenta. A map of update constraints in inductive inference. In *Proc. of ALT (Algorithmic Learning Theory)*, 2014.
18 S. Lange and T. Zeugmann. Monotonic versus non-monotonic language learning. In *Proc. of Nonmonotonic and Inductive Logic*, pages 254–269, 1993.
19 D. Osherson, M. Stob, and S. Weinstein. Learning strategies. *Information and Control*, 53:32–51, 1982.
20 D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*. MIT Press, Cambridge, MA, 1986.
21 H. Rogers. *Theory of Recursive Functions and Effective Computability*. McGraw Hill, New York City, NY, 1967. Reprinted by MIT Press, Cambridge, MA, 1987.
22 G. Schäfer-Richter. *Über Eingabeabhängigkeit und Komplexität von Inferenzstrategien*. PhD thesis, RWTH Aachen, 1984.
23 K. Wexler and P. Culicover. *Formal Principles of Language Acquisition*. MIT Press, Cambridge, MA, 1980.
24 R. Wiehagen. A thesis in inductive inference. In *Proc. of Nonmonotonic and Inductive Logic*, pages 184–207, 1991.